ISSN 1870-4069

Desempeño de algoritmos de inteligencia artificial en la clasificación de objetos astronómicos en Gaia DR3

Orestes Javier Pérez-Cruz[1], Cynthia Alejandra Martínez-Pinto[1], Silvana Guadalupe Navarro-Jiménez[2], Luis José Corral-Escobedo[2], Marco Antonio Meza Aguilar[1]

[1] Instituto Tecnológico de Ciudad Guzmán, Tecnológico Nacional de México, México

[2] Universidad de Guadalajara, Instituto de Astronomía y Meteorología, México

{orestesperez1995, silvananj}@gmail.com, cynthia_amp@hotmail.com, luis.corral@academicos.udg.mx, marco.ma@cdguzman.tecnm.mx

Resumen. Se realiza un análisis del desempeño de una serie de algoritmos de Machine Learning (ML) para la clasificación de objetos astronómicos utilizando los datos del catálogo DR3 de la misión espacial Gaia. Los modelos de aprendizaje automático fueron entrenados con la información espectral proveniente de los espectrofotómetros rojo y azul del satélite. El propósito es lograr una clasificación precisa de los siguientes tipos de objetos: Estrellas Simbióticas, Nebulosas Planetarias y Gigantes Rojas. Se evalúan diferentes algoritmos de clasificación, incluyendo Random Forest (RF), Máquina de Soporte Vectorial (SVM), Redes Neuronales Artificiales (RNA) y Gradient Boosting. Se comparan los resultados obtenidos usando diversas métricas (Precision, Recall, F1-Score, el índice Kappa) y se comprueba la efectividad de clasificar las estrellas antes mencionadas, utilizando solamente la información de sus espectros. Los modelos que obtuvieron los mejores resultados fueron los entrenados con Redes Neuronales Artificiales y Random Forest, con un porcentaje de precisión superior al 94.67%.

Palabras claves: Clasificación automática, machine learning, Gaia DR3, espectroscopia, objetos astronómicos, nebulosas planetarias, estrellas simbióticas, gigantes rojos.

Performance of Artificial Intelligence Algorithms in the Classification of Astronomical Objects in Gaia DR3

Abstract. An analysis is conducted on the performance of a series of Machine Learning (ML) algorithms for the classification of astronomical objects using

data from the DR3 catalog of the Gaia space mission. The machine learning models were trained using spectral information from the satellite's red and blue spectrophotometers. The purpose is to achieve accurate classification of the following types of objects: Symbiotic Stars, Planetary Nebulae, and Red Giants. Different classification algorithms, including Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Gradient Boosting, are evaluated. The results are compared using various metrics such as Precision, Recall, F1-Score, and the Kappa index. The effectiveness of classifying the stars using only their spectral information is demonstrated. The models trained with Artificial Neural Networks and Random Forest achieved the best results, with a precision percentage exceeding 94.67%.

Keywords: Automatic classification, machine learning, Gaia DR3, spectroscopy, astronomical objects, planetary nebulae, symbiotic stars, red giants.

1. Introducción

El catálogo Gaia DR3 (GDR3), liberado el 13 de junio de 2022, incluye por primera vez, los espectros calibrados de los objetos astronómicos: uno que cubre la región azul del espectro electromagnético (BP) abarcando longitudes de onda de 330 a 680 nm, y otro en la región roja (RP) que cubre el rango de 640 a 1050 nm [1]. Los espectros medios BP/RP observados son de baja resolución, llegando a magnitudes de $G < 17.65$ [2].

Dentro del catálogo, existen estrellas con características peculiares que resultan difíciles de clasificar utilizando los métodos convencionales. Además, el enorme volumen de información que un telescopio moderno puede generar impide a los astrónomos procesar dichos datos de manera individual. Por lo tanto, la clasificación automática es un imperativo en los tiempos actuales, donde se manejan una gran cantidad de datos. Gaia, en su catálogo DR3, liberó alrededor de 470 millones de fuentes con parámetros astrofísicos de espectros BP/RP [3].

La presente investigación se enfoca en tres tipos de objetos astronómicos: estrellas simbióticas, nebulosas planetarias y gigantes rojas. Estas estrellas en particular son de interés debido a que provienen de la evolución de cuerpos celestes de baja masa, y debido a la similitud de sus espectros, provoca que, en ocasiones, se clasifiquen erróneamente.

– **Gigante Roja (RG).** Son estrellas de masa baja e intermedia (entre 0.8 y 8 Mo) que han evolucionado. A medida que envejecen, van agotando el Hidrógeno que sostiene la fusión nuclear que ocurre en el centro de la estrella. Esto hace que el núcleo se contraiga, aumentando la temperatura y la presión, provocando la expansión de las capas exteriores. Este proceso culminará en la formación de una enana blanca, que representa la etapa final del ciclo de vida de estas estrellas. Al liberar una gran cantidad de elementos producto de la fusión nuclear al medio interestelar, las enanas blancas desempeñan un papel fundamental en la evolución química de las galaxias en las que se encuentran [4].

– **Estrellas simbióticas (SS)**. Son sistemas estelares compuestos por dos estrellas separadas que orbitan entre sí. Estos sistemas están constituidos por una gigante roja evolucionada (tipo espectral K o M) que pierde y transfiere masa a su segundo componente. Por lo general, esta segunda componente es una enana blanca, caracterizada por su alta temperatura y emisión de una gran cantidad de fotones ionizantes. Se habla de simbiosis estelar debido a que cada una de las estrellas depende e influye en la evolución de la otra. La comprensión de los procesos de transferencia de masa y de acreción en estos sistemas es importante para entender la evolución de las estrellas y cualquier interacción binaria que involucre a gigantes evolucionados [5].

– **Nebulosa Planetaria (PN).** Es una nube circunestelar ionizada en expansión que fue expulsada durante la fase de la rama asintótica gigante (GRB) de su estrella progenitora, una estrella por debajo de 8 o 9 masas solares [6]. De la estrella queda un residuo en forma de enana blanca, el cual se encuentra a una gran temperatura. Estas nebulosas, en general, forman anillos o burbujas, pero, debido a las características del material circundante o al carácter binario del progenitor (como en el caso de las simbióticas), pueden ser también elipsoidales, bipolares o hasta cuadrupolares [7].

De los objetos astronómicos antes mencionados, las PNs y SS son difíciles de distinguir entre ellas, debido a las características que poseen y su baja representación con respecto a otros objetos celestes. Estas estrellas se caracterizan por presentar intensas líneas de emisión en su espectro visible.

Debido al volumen de estos objetos, se requiere el uso de algoritmos de aprendizaje automático (ML) para su clasificación automatizada. Después de examinar investigaciones anteriores, se decidió emplear los modelos que han demostrado los mejores resultados, como: Random Forest (RF), Máquina de Soporte Vectorial (SVM), Redes Neuronales Artificiales (RNA) y Gradient Boosting. Estos modelos se entrenaron utilizando la información espectral del catálogo Gaia DR3.

2. Metodología

2.1. Adquisición y tratamiento de los datos

En este estudio, se empleó la información suministrada por SIMBAD[1] para identificar las estrellas analizadas. SIMBAD se destaca por su capacidad de ofrecer información detallada sobre objetos astronómicos presentes en artículos científicos [8].

Posteriormente se procedió a realizar una búsqueda cruzada en la tabla *xp_continuous_mean_spectrum* de la base de datos de Gaia DR3 para determinar los objetos astronómicos por tipo de estrella. Esta tabla proporciona la media de los espectros BP y RP basados en una representación continua en funciones base [9].

Para evitar posibles pérdidas de información al muestrear los espectros, los espectros calibrados se encuentran representados como una combinación lineal de funciones base en lugar de utilizar la convencional tabla de flujos y longitudes de onda [10].

[1] SIMBAD. Base de datos dinámica de objetos astronómicos (https://simbad.cds.unistra.fr/simbad/)

Los espectros descargados originalmente se encontraban internamente calibrados en los rangos de longitud de onda BP y RP. Para su procesamiento, se utilizó la librería GaiaXPy[2], la cual permitió calibrar y muestrear cada espectro en una cuadrícula de longitud de onda uniforme predeterminada mediante la rutina *calibrate*. De esta manera, se obtuvo un único espectro que abarcaba el rango completo de longitud de onda cubierto por BP y RP. El proceso de calibración anterior generó un total de 343 valores de flujo por espectro. Se utilizó un muestreo predeterminado que abarcó un rango de longitud de onda de 336 a 1020 nm, con un incremento de 2 nm entre cada punto de muestreo.

En total, se obtuvieron los siguientes recuentos de espectros por tipo de estrellas del catálogo Gaia DR3, disponible en el sitio Gaia Archive[3]: 201 espectros correspondientes a estrellas Simbióticas, 574 a Nebulosas Planetarias y 69,146 a Gigantes Rojas. Es notable que el número de espectros de Gigantes Rojas es significativamente superior en comparación con los otros tipos de estrellas, lo cual se debe a su mayor presencia en el universo. Con el fin de obtener una distribución más uniforme, se seleccionó solamente una muestra de 1200 espectros de Gigantes Rojas para su inclusión en este estudio.

2.2. Preprocesamiento de los datos

Con el objetivo de mejorar el rendimiento y la estabilidad de los algoritmos de aprendizaje automático durante el entrenamiento y la inferencia, se normalizaron los valores de flujo de cada espectro en una escala de 0 a 1. De esta manera, todos los valores del espectro quedaron expresados como una fracción del valor máximo, lo que permitió establecer una escala común entre los distintos espectros (Fig. 1).

El proceso de normalización permite, aún a simple vista, distinguir mejor entre los diferentes tipos de espectros. El espectro de la PN consta casi exclusivamente de líneas de emisión, en cambio las RG poseen principalmente un continuo con gran número de líneas y/o bandas de absorción. Por su parte las SS resultan una combinación de ambos tipos de espectros, mostrando líneas en emisión y bandas en absorción en la región IR del espectro (700 a 1000 nm).

Posterior a la normalización, se conformó el conjunto de datos que se utilizaría como entrada para los algoritmos de ML. Este conjunto de datos consta de 1,975 registros que representan los espectros de las estrellas de interés. Cada registro está compuesto por 343 características, que corresponden a los valores de flujo normalizados en el rango de longitud de onda de 336 a 1020 nm. Además, se incluyó una columna adicional en el conjunto de datos con la etiqueta correspondiente al tipo de estrella.

Como se puede evidenciar los datos presentan un desbalanceo notable, debido a que existe una diferencia significativa en la cantidad de muestras por cada tipo de estrella. El desbalanceo de datos puede tener un impacto negativo en el rendimiento de los algoritmos de aprendizaje automático, debido a que pueden presentar dificultades para

[2] GaiaXPy: Biblioteca Python para el análisis de datos astronómicos de la misión Gaia (https://gaia-dpci.github.io/GaiaXPy-website/)

[3] Gaia Archive: Repositorio en línea de datos astronómicos de la misión espacial Gaia (https://gea.esac.esa.int/archive/)

Tabla 1. Número de estrellas que conforman los conjuntos de datos desbalanceados y balanceados respectivamente.

Estrellas	Conjunto de Datos	
	Datos desbalanceados	**Datos balanceados**
Estrellas Simbióticas	201	1000
Nebulosas Planetarias	574	1000
Gigantes Rojas	1200	1000

Tabla 2. Descripción de los parámetros utilizados en el entrenamiento del algoritmo Random Forest.

Parámetro	Selección
Número de árboles	[10, 500]
Profundidad máxima	Sin restricción
Criterio	gini (impureza de Gini)
Máximo de características	Total de características
Número máximo de hojas	Sin límite

Tabla 3. Descripción de los parámetros utilizados en el entrenamiento del algoritmo SVM.

Parámetro	Selección
Kernel	['linear', 'poly', 'rbf']
C (Parámetro de regularización)	1.0
Grado usando el kernel polinómico	3

aprender patrones y tomar decisiones precisas para las clases minoritarias. Esta afirmación se pudo comprobar en la sección de resultados del estudio.

En consecuencia, se llevó a cabo la creación de un conjunto de datos balanceado, garantizando que cada tipo de estrella estuviera representado por 1,000 muestras.

En el caso de las Gigantes Rojas, no se encontraron dificultades debido a que la cantidad recuperada superaba esta cifra. Por lo tanto, se seleccionaron muestras aleatoriamente hasta obtener la cantidad deseada.

Sin embargo, en el caso de las estrellas simbióticas y las nebulosas planetarias, la cantidad de muestras era insuficiente, por lo que se decidió generar nuevos espectros a partir de los originales. Para lograr esto, se empleó el método de adición de ruido blanco, en el cual se generó una secuencia de números aleatorios siguiendo una distribución normal con media 0 y desviación estándar variable, en este caso, en el rango de 0.01 a 0.05. El proceso de generación de nuevos espectros implicó combinar los datos originales con el ruido blanco generado (Ver Fig. 2).

Este nuevo conjunto de datos balanceados, al igual que el anterior estaría conformado por 343 características que representan los valores de flujo y la columna adicional que representa la etiqueta del espectro. En este caso, se logró alcanzar una cantidad final de 3,000 espectros en total, distribuidos equitativamente con 1,000 espectros por cada tipo de estrella. Este conjunto de datos equilibrado garantiza que cada tipo de estrella esté representado de manera adecuada y proporcional en el conjunto. Sin embargo, es importante tener precaución al interpretar los resultados

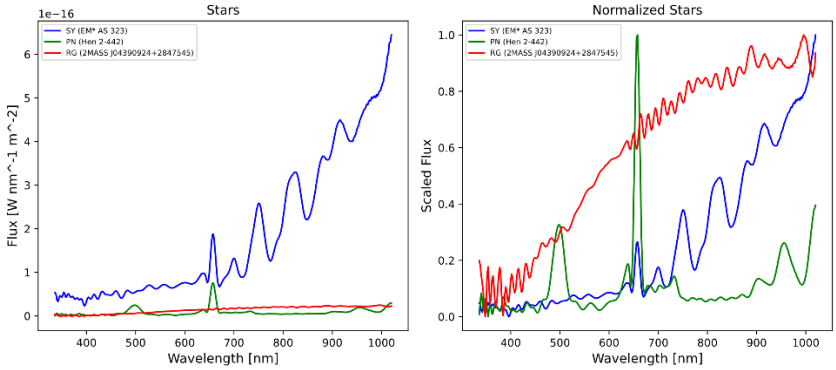

Fig. 1. Las gráficas corresponden a tres tipos de estrellas diferentes, *(azul)* Estrella simbiótica [EM* AS 323], *(rojo)* Gigante Roja [2MASS J04390924+2847545], *(verde)* Nebulosa Planetaria [Hen 2-442]. En la primera gráfica los valores de flujo están expresados en Vatios por nanómetro por metro cuadrado (W/nm/m²), resultado de calibrarlos externamente usando la librería GaiaXpy. En la gráfica de la derecha los valores de flujo se encuentran escalados en un rango de 0 a 1.

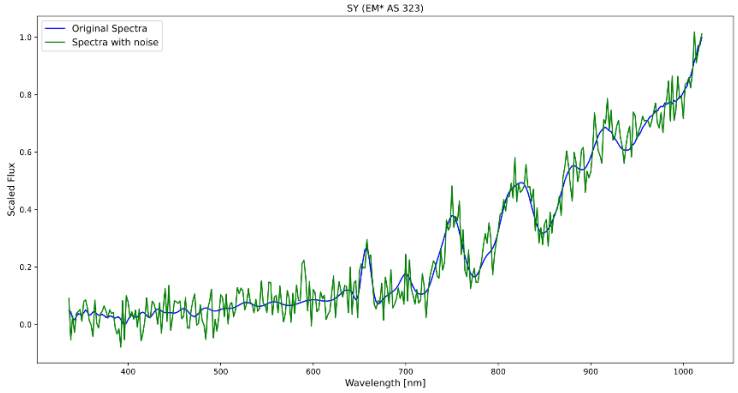

Fig. 2. En la gráfica se muestra al espectro de la estrella simbiótica original *(azul)*. Y al espectro resultado de la adición de ruido *(verde)*, siguiendo una distribución normal con media 0 y desviación estándar 0.05.

obtenidos mediante el sobremuestreo. Los ejemplos sintéticos generados pueden introducir sesgos y afectar la capacidad del modelo de generalización en nuevos datos [11]. La Tabla 1 muestra como quedaron conformados los conjuntos de datos resultantes.

2.3. Análisis y selección de los algoritmos

Cada uno de los dos conjuntos de datos creados fue dividido en dos subconjuntos. El primer subconjunto, con una representación del 80% del total de muestras, se utilizó para entrenar los diferentes algoritmos de ML. El otro subconjunto, que representaba

el 20% restante, se reservó para las pruebas. Este conjunto de datos se empleó exclusivamente para la evaluación, con el objetivo de determinar si los algoritmos fueron capaces de aprender y generalizar adecuadamente sin llegar a un sobreajuste.

Para el análisis se utilizaron los siguientes algoritmos supervisados de ML para clasificación: Random Forest, Máquina de Soporte Vectorial, Redes Neuronales Artificiales y Gradient Boosting. La elección de estos algoritmos proporciona una diversa combinación de enfoques de clasificación, lo que permite evaluar su rendimiento y comparar sus resultados en el conjunto de pruebas, para determinar cuál de ellos se adapta mejor a nuestro problema.

Random Forest

Random Forest es un algoritmo de aprendizaje automático supervisado que combina múltiples árboles predictores. Cada árbol se construye basándose en un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque [12]. Los árboles de decisión tienden a sobreajustarse, lo que significa que aprenden con precisión los datos de entrenamiento, pero tienen dificultades para aplicar ese conocimiento a nuevos datos.

La capacidad de generalización del algoritmo puede mejorarse al combinar múltiples árboles en un conjunto, utilizando la técnica conocida como *ensemble*. Sin embargo, al aplicar Random Forest, es importante tener en cuenta que puede aumentar la complejidad de la interpretación del modelo. Además, este algoritmo es sensible a datos altamente correlacionados, lo cual puede disminuir su capacidad de generalización.

Se realizaron varias pruebas de entrenamiento, cambiando los parámetros que recibía el algoritmo en cada una (ver Tabla 1).).

Máquina de soporte vectorial

La Máquina de Soporte Vectorial (SVM) es un algoritmo de aprendizaje automático supervisado que se utiliza principalmente para la clasificación de datos. En lugar de trabajar directamente en los datos originales, la SVM los representa como puntos en un espacio de múltiples dimensiones, lo que facilita la visualización y el análisis de las relaciones entre las variables. Sin embargo, la interpretación de las decisiones de clasificación es limitada, ya que se enfoca en encontrar un hiperplano óptimo de separación [13]. La comprensión completa del modelo y las relaciones entre las características se dificulta debido a que la atención se centra en los vectores de soporte, que representan solo una pequeña fracción del conjunto de datos total.

Se realizaron pruebas con diferentes parámetros, utilizando distintos *kernels* en cada uno. En la Tabla 3 se muestran las configuraciones analizadas.

Redes Neuronales Artificiales

Las redes neuronales artificiales (ANN) son un subconjunto de herramientas de aprendizaje automático y forman parte de los algoritmos de aprendizaje profundo (deep learning). Su nombre y estructura están inspirados en el cerebro humano, imitando la forma en que las neuronas biológicas envían señales entre sí. Poseen altas velocidades de procesamiento y la capacidad de aprender la solución a un problema a partir de un conjunto de ejemplo [14]. Sin embargo, es importante considerar las limitaciones en la interpretación de las redes neuronales artificiales. Estas limitaciones abarcan su

Tabla 4. Descripción de los parámetros utilizados en el entrenamiento del algoritmo Redes Neuronales Artificiales.

Parámetro	Selección
Capas	(343 - 64 - 32 - 32 - 3)
Función de activación	ReLU
Función de perdida	sparse_categorical_crossentropy
Optimizador	Adam

Tabla 5. Descripción de los parámetros utilizados en el entrenamiento del algoritmo Gradient Boosting.

Parámetro	Selección
Función de pérdida	('log_loss', 'deviance', ' exponential')
Cantidad de estimadores	[100-500]
Taza de aprendizaje	[0.1-0.5]
Profundidad máxima	3
Cantidad de características	Total de características

naturaleza de “caja negra”, su sensibilidad a los datos de entrada, el riesgo de sobreajuste y la dificultad de interpretar características abstractas aprendidas.

La red neuronal diseñada tiene la siguiente topología: una capa de entrada de 343 neuronas, seguida de una capa oculta de 64 neuronas y dos capas ocultas de 32 neuronas cada una. Todas las capas son densas, lo que significa que todas las neuronas están completamente conectadas, además utilizan la función de activación *ReLU* para introducir no linealidad en los datos. Después de cada capa densa, se añade una capa *Dropout*, que desactiva aleatoriamente el 10% de las neuronas durante el entrenamiento. Esto ayuda a prevenir el sobreajuste (overfitting) y mejora la capacidad de generalización del modelo. La capa de salida consta de 3 neuronas y utiliza la función de activación *softmax*, comúnmente empleada en problemas de clasificación multiclase. En la tabla 4 se muestra la configuración de la red neuronal.

Gradient Boosting

Gradient Boosting es un algoritmo que se enfoca en la optimización numérica del espacio de funciones en lugar del espacio de parámetros. Trabaja de manera iterativa, donde en cada etapa se añade un nuevo componente a la aproximación existente, ajustándolo en función del gradiente de la función de pérdida. Esto permite mejorar gradualmente la aproximación y obtener mejores resultados tanto en problemas de regresión como de clasificación [15]. Sin embargo, este algoritmo puede presentar limitaciones en términos de interpretabilidad. Estas incluyen la complejidad del modelo, la captura de relaciones no lineales, la dificultad para identificar interacciones entre variables y las características de alta dimensionalidad.

Se probaron diferentes combinaciones de parámetros, utilizando distintas funciones de pérdida, diferentes valores de taza de aprendizaje, entre otros, quedando las siguientes configuraciones (Ver Tabla 5).

Matriz de Confusión [Random Forest]

Datos desbalanceados

$$\begin{bmatrix} 30 & 6 & 4 \\ 4 & 97 & 14 \\ 0 & 5 & 235 \end{bmatrix}$$

Datos balanceados

$$\begin{bmatrix} 191 & 5 & 4 \\ 5 & 187 & 8 \\ 4 & 6 & 190 \end{bmatrix}$$

Fig. 3. Matriz de confusión resultado de la ejecución del algoritmo Random Forest.

Tabla 6. Métricas obtenidas del algoritmo de Random Forest.

Métricas	Datos desbalanceados	Datos balanceados
Precision	0.9031	0.9467
Recall	0.8575	0.9467
F1-Score	0.8780	0.9467
Kappa	0.8401	0.9200

3. Resultados

3.1. Definición de métricas

Para comparar la precisión de los algoritmos presentados anteriormente, se utilizaron las siguientes métricas: Precisión (Precision), Puntuación F1 (F1-score), Recall (Sensibilidad) y el coeficiente Kappa de Cohen. Estas métricas se calcularon luego de que los algoritmos evaluaran el conjunto de datos de prueba utilizando la matriz de confusión. Esta matriz muestra el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) para cada clase.

F1-score es una medida que combina la Precisión y el Recall en un solo valor. Proporciona una medida equilibrada entre la precisión y la capacidad de recuperación del clasificador [16]. Es especialmente útil cuando el conjunto de datos está desequilibrado en términos de clases. La fórmula es la siguiente:

$$F1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}.$$

El coeficiente Kappa de Cohen, es una medida que expresa el nivel de acuerdo entre dos anotadores en un problema de clasificación [17]. Se define de la siguiente manera:

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}.$$

En nuestro estudio de clasificación multiclase, se adaptaron las métricas de evaluación utilizadas en problemas de clasificación binaria para su aplicación en un contexto multiclase. Para evaluar el rendimiento de nuestro modelo de clasificación se utilizó la técnica *macro-averaged* [18].

Matriz de Confusión [SVM]

Datos desbalanceados

$$\begin{bmatrix} 22 & 7 & 11 \\ 4 & 93 & 18 \\ 3 & 1 & 236 \end{bmatrix}$$

Datos balanceados

$$\begin{bmatrix} 189 & 9 & 2 \\ 13 & 180 & 7 \\ 3 & 3 & 194 \end{bmatrix}$$

Fig. 4. Matriz de confusión resultado de la ejecución del algoritmo SVM.

Tabla 7. Métricas obtenidas del algoritmo SVM.

Métricas	Datos desbalanceados	Datos balanceados
Precision	0.8567	0.9384
Recall	0.7807	0.9383
F1-Score	0.8111	0.9382
Kappa	0.7818	0.9075

La elección de este enfoque se basó en la necesidad de tratar todas las clases de manera equitativa durante la evaluación, independientemente de su tamaño o distribución de los datos. En primer lugar, se calcularon las métricas de manera binaria para cada clase de forma individual, y posteriormente se obtuvo el promedio de estas métricas para obtener una evaluación global del modelo.

3.2. Modelos

Random Forest

La Fig. muestra la matriz de confusión generada al ejecutar el algoritmo Random Forest. Se observa que la clasificación de las estrellas simbióticas y las nebulosas planetarias tuvieron una mayor imprecisión. Sin embargo, en el conjunto de datos balanceados, a pesar de la presencia de errores, estos son menores, en comparación con el conjunto de datos desbalanceados. Esto se refleja en el valor de Recall, que mejoró de 0.8575 a 0.9467, indicando una mayor capacidad del modelo para identificar correctamente las estrellas. La Tabla 6 muestra los valores obtenidos de todas las métricas evaluadas usando el algoritmo Random Forest.

Máquina de Soporte Vectorial

El algoritmo SVM mostró un rendimiento deficiente en las pruebas realizadas con el conjunto de datos desbalanceados. En este escenario, el modelo tendió a sobreajustarse y no logró generalizar correctamente, lo que resultó en una clasificación incorrecta del 12 % del total de muestras usando el conjunto de datos desbalanceados (ver Fig. 4). No obstante, al utilizar el conjunto de datos balanceados, se observó una mejora significativa en el rendimiento, ya que se redujo considerablemente el error de clasificación (6.16%), pasando de un valor de sensibilidad de 0.7807 a 0.9383. La Tabla 7 muestra los valores obtenidos de todas las métricas evaluadas usando el algoritmo SVM.

Matriz de Confusión [Gradient Boosting]

Datos desbalanceados

$$\begin{bmatrix} 31 & 3 & 6 \\ 6 & 93 & 16 \\ 0 & 7 & 233 \end{bmatrix}$$

Datos balanceados

$$\begin{bmatrix} 186 & 9 & 5 \\ 8 & 180 & 12 \\ 1 & 8 & 191 \end{bmatrix}$$

Fig. 5. Matriz de confusión resultado de la ejecución del algoritmo Gradient Boosting.

Tabla 8. Métricas obtenidas del algoritmo Gradient Boosting.

Métricas	Datos desbalanceados	Datos balanceados
Precision	0.8848	0.9286
Recall	0.8515	0.9283
F1-Score	0.8666	0.9283
Kappa	0.8158	0.8925

Gradient Boosting

El algoritmo Gradient Boosting demostró un rendimiento similar al obtenido por Random Forest en nuestro estudio. Sin embargo, existe la presencia de falsos positivos, debido al desequilibrio presente en los datos de entrenamiento. No obstante, se logra obtener mejoras significativas al emplear el conjunto de datos balanceado (Ver Fig. 5). El valor de F1-Score aumentó de 0.8666 a 0.9283. Este incremento indica una mayor capacidad del algoritmo para identificar correctamente las estrellas en cuestión, reduciendo así los falsos positivos. La Tabla 8 muestra los valores obtenidos de todas las métricas evaluadas usando el algoritmo Gradient Boosting.

Redes Neuronales Artificiales

El algoritmo de Redes Neuronales Artificiales mostró resultados superiores en comparación con los algoritmos anteriores, logrando los valores más altos de precisión en general. Esto se evidencia en la matriz de confusión mostrada en la Fig. 6, donde se observa una disminución en el número de falsos positivos tanto en el conjunto de datos balanceados como en los desbalanceados. El valor de F1-Score respalda esta afirmación, con un valor de 0.9012 y 0.9533 para el conjunto de datos desbalanceados y balanceados respectivamente (Ver Tabla 9). Su capacidad para capturar relaciones complejas entre los datos y adaptarse a diferentes patrones lo posiciona como una opción favorable, especialmente en casos de conjuntos de datos desbalanceados.

4. Conclusiones

El estudio se centró en la clasificación de los siguientes objetos astronómicos: Estrellas Simbióticas, Nebulosas Planetarias y Gigantes Rojas. Se observó que las Gigantes Rojas eran más numerosas y representaban una proporción mayor en comparación con las restantes clases. Esta desigualdad generó dificultades en la clasificación, ya que los modelos tendieron a confundir parte de la clase minoritaria

Matriz de Confusión [RNA]

Datos desbalanceados

$$\begin{bmatrix} 33 & 4 & 3 \\ 2 & 97 & 16 \\ 1 & 1 & 238 \end{bmatrix}$$

Datos balanceados

$$\begin{bmatrix} 197 & 2 & 1 \\ 5 & 186 & 9 \\ 1 & 10 & 189 \end{bmatrix}$$

Fig. 6. Matriz de confusión resultado de la ejecución del algoritmo RNA.

Tabla 9. Métricas obtenidas del algoritmo RNA.

Métricas	Datos desbalanceados	Datos balanceados
Precision	0.9312	0.9532
Recall	0.8867	0.9533
F1-Score	0.9067	0.9532
Kappa	0.8686	0.9300

(Estrellas Simbióticas y Nebulosas Planetarias) con la clase mayoritaria (Gigantes Rojas).

Se pudo comprobar que los algoritmos Random Forest y Redes Neuronales Artificiales mostraron resultados satisfactorios, con valores de precisión de 0.9467 y 0.9532, respectivamente, para el conjunto de datos balanceado. Por otro lado, en el conjunto de datos desbalanceado, Random Forest alcanzó una precisión de 0.9031, mientras que la RNA alcanzó una precisión de 0.9312. Por lo tanto, estos algoritmos demuestran una mayor efectividad en conjuntos de datos desbalanceados en comparación con otros enfoques.

Las Redes Neuronales Artificiales (RNA) demostraron ser altamente efectivas al trabajar con conjuntos de datos desbalanceados debido a su capacidad para aprender patrones complejos y adaptarse a diferentes distribuciones de clases.

Estos modelos de clasificación se presentan como una herramienta valiosa y de gran utilidad para los astrónomos, al brindarles un apoyo efectivo en la clasificación de estrellas peculiares. Su aplicación contribuye significativamente a una mejor comprensión del ciclo de vida de las estrellas.

5. Trabajo futuro

Como trabajo futuro, se sugiere realizar pruebas y evaluaciones de estos modelos de clasificación utilizando conjuntos de datos de estrellas candidatas a Estrellas Simbióticas y Nebulosas Planetarias. Esto permitiría comprobar la capacidad de los modelos para identificar y clasificar de manera precisa estos tipos de estrellas.

Además, se propone la expansión de los modelos para incluir y clasificar una mayor variedad de tipos de estrellas, como: Estrellas Be, Variables Cataclísmicas, Estrellas Mira, Estrellas AeBe, Estrellas post-AGB, Estrellas K-gigantes, entre otras.

Esta ampliación permitirá una aplicación más versátil del modelo, facilitando la identificación de estrellas que no se encuentran clasificadas dentro del catálogo de Gaia.

Referencias

1. Prusti, T., De Bruijne, J.H.J., Brown, A.G., Vallenari, A., Babusiaux, C., Bailer-Jones, C.A.L., and Navascués, D.B.: The gaia mission. Astronomy and Astrophysics, vol. 595, pp. 36 (2016). DOI: 10.1051/0004-6361/201629272.
2. Vallenari, A., Brown, AG., Prusti, T., De Bruijne, J.H., Arenou, F., Babusiaux, C., and Bianchi, L.: Gaia data release 3-summary of the content and survey properties. Astronomy and Astrophysics, vol. 674 (2022). DOI: 10.1051/0004-6361/202243940.
3. Gaia Data Release 3 contents summary-Gaia-Cosmos: https://www.cosmos.esa.int/web/gaia/dr3 (2022)
4. Jastrow, R.: Red Giants and White Dwarfs. vol. 269 (1990)
5. Mikolajewska, J.: Symbiotic Stars: Observations confront theory. Baltic Astronomy, vol. 21, pp. 5–12 (2012). DOI: 10.1515/ASTRO-2017-0352.
6. Frankowski, A., Soker, N.: Very late thermal pulses influenced by accretion in planetary nebulae. New Astronomy, vol. 14, no. 8, pp. 654–658 (2009). DOI: 10.1016/J.NEWAST.2009.03.006.
7. Kwok, S.: The origin and evolution of planetary nebulae. The Origin and Evolution of Planetary Nebulae/Sun Kwok, Cambridge University Press (2000)
8. Wenger, M., Ochsenbein, F., Egret, D., Dubois, P., Bonnarel, F., Borde, S., and Monier, R.: The SIMBAD Astronomical Database-The CDS Reference Database for Astronomical objects. Astronomy and Astrophysics Supplement Series, vol. 143, no. 1, pp. 9–22 (2000). DOI: 10.1051/AAS:2000332.
9. Babusiaux, C., Fabricius, C., Khanna, S., Muraveva, T., Reylé, C., Spoto, F., and Weiler, M.: Gaia Data Release 3-Catalogue validation. Astronomy and Astrophysics, vol. 674, no. A32, pp. 1–25 (2023). DOI: 10.1051/0004-6361/202243790.
10. Carrasco, J.M., Weiler, M., Jordi, C., Fabricius, C., De Angeli, F., Evans, D.W., and Montegriffo, P.: Internal Calibration of Gaia BP/RP Low-Resolution Spectra. Astronomy and Astrophysics, vol. 652, no. A86 (2021). DOI: 10.1051/0004-6361/202141249.
11. Santos, M.S., Soares, J.P., Abreu, P.H., Araujo, H., and Santos, J.: Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [research frontier]. In: IEEE Computational Intelligence Magazine, vol. 13, no. 4, pp. 59–76 (2018). DOI: 10.1109/MCI.2018.2866730.
12. Breiman, L.: Random forests. Machine Learning 2001, vol. 45, pp. 5–32 (2001). DOI: 10.1023/A:1010933404324.
13. Noble, W.S.: What is a support vector machine?. Nature Biotechnology, vol. 24, no. 12, pp. 1565–1567 (2006). DOI: 10.1038/nbt1206-1565.
14. Bishop, C.M.: Neural Networks and their Applications. Review of scientific instruments, vol. 65, no. 6, pp. 1803–1832 (1998). DOI: 10.1063/1.1144830.
15. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of statistics, vol. 29, no. 4, pp. 1189–1232, (2001). DOI: 10.1214/AOS/1013203451.
16. Powers, D.M.: Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. arXiv:2010.16061 (2020)
17. Chicco, D., Warrens, M.J., and Jurman, G.: The Matthews Correlation Coefficient (MCC) is more Informative than Cohen's Kappa and Brier Score in Binary Classification Assessment. IEEE, vol. 9, pp. 78368–78381 (2021). DOI: 10.1109/ACCESS.2021.3084050.
18. Farhadpour, S., Warner, T.A., and Maxwell, A.E.: Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and best Practices. Remote Sensing, vol. 16, no. 3, pp. 533 (2024). DOI: 10.3390/RS16030533.